Indice

Capitolo 1

Introduzione

Nel contesto delle variabili quantitative, esistono tecniche di analisi multivariata che permettono di ridurre la dimensione di un fenomeno mantenendo un elevato grado di informazione.

Tecniche come l'analisi delle componenti principali, l'analisi della correlazione canonica e l'analisi fattoriale consentono, in ambito metrico, di creare strutture semplificate per elaborazioni successive garantendo, in questo modo, il principio di parsimonia.

Invece, l'impossibilità di attribuire valori numerici alle modalità di variabili di natura qualitativa, ha rallentato notevolmente la ricerca di analoghi metodi per l'analisi multivariata di dati categoriali. Per ovviare a tale inconveniente si è tentato, in vari modi, di adattare i modelli di analisi multivariata per variabili metriche a variabili qualitative, non sempre con risultati soddisfacenti.

In questa tesi, prendendo spunto dal questionario sulla valutazione della didattica universitaria nell'Ateneo di Siena (A.A. 1998/1999), si tenterà di adattare il metodo di analisi delle componenti principali a variabili ordinali, senza modificare la natura qualitativa dei dati.

La tesi presenta, inizialmente, la descrizione degli aspetti teorici dell'adattamento del metodo delle componenti principali a variabili ordinali e, successivamente, l'applicazione di tale metodo a insiemi di dati reali: i risultati di un'indagine sulla

valutazione della didattica nell'Ateneo di Siena, con riferimento sia ai dati globali dell'Ateneo che della Facoltà di Economia dello stesso Ateneo.

In particolare, nelle prime due sezioni del secondo capitolo, saranno brevemente richiamati il metodo delle componenti principali e alcune tra le più importanti misure di associazione ordinale; nelle sezioni successive, si esamineranno le condizioni di applicabilità del metodo a variabili ordinali:

- il momento primo ed il momento secondo dovranno caratterizzare totalmente la distribuzione multivariata, in quanto solo in questo caso le componenti principali potranno fornire informazioni complete sul fenomeno studiato e l'utilizzo del metodo sarà pienamente giustificato;
- la matrice di correlazione sarà sostituita con la matrice di associazione e, mediante questa operazione, i dati manterranno la loro natura qualitativa;
- la matrice di associazione dovrà soddisfare le proprietà di una matrice di correlazione per poter essere utilizzata al suo posto all'interno della tecnica delle componenti principali.

L'ultimo paragrafo del medesimo capitolo metterà in evidenza le potenzialità del metodo nella formazione di graduatorie, che rappresenta uno degli impieghi tradizionali del metodo delle componenti principali. Se, tra le componenti individuate, la prima spiega un'alta porzione di variabilità totale, essa potrà essere utilizzata per la formazione di graduatorie.

Nel terzo capitolo, nell'applicazione ai dati sulla valutazione della didattica nell'Ateneo di Siena si evidenzierà un caso di inapplicabilità del metodo; nell'applicazione ai dati relativi alla Facoltà di Economia dello stesso Ateneo, invece, sarà possibile procedere con la semplificazione della struttura dei dati fino alla formazione di graduatorie dei diversi corsi della Facoltà considerata.

Capitolo 2

Componenti principali ordinali

L'analisi delle componenti principali, così come l'analisi della correlazione canonica e l'analisi fattoriale, è un metodo di analisi multivariata utilizzato per ridurre la dimensionalità di dati multivariati ed ottenere una sintesi delle informazioni più agevole.

Lavorando su variabili ordinali, il percorso più intuitivo da seguire per utilizzare il metodo delle componenti principali sarebbe quello di attribuire punteggi numerici alle modalità di risposta e trattare le variabili come numeriche. Ciò presuppone, però, la conoscenza delle distanze tra le categorie e, dal momento che, in generale, è noto solo l'ordinamento tra una modalità di risposta e l'altra, questo procedimento appare spesso una forzatura.

E' possibile, invece, mantenere la natura ordinale dei dati quando l'attribuzione di etichette numeriche ha il solo scopo di rappresentare l'ordinamento delle modalità. Quando si vuole conoscere, ad esempio, il parere dei dipendenti di un'azienda in merito al trattamento economico loro riservato, ci si può riferire alle seguenti categorie:

1=per niente soddisfatto

2=poco soddisfatto

3=mediamente soddisfatto

4=soddisfatto

5=molto soddisfatto.

Se a tali punteggi si attribuisce un significato di equidistanza tra due categorie contigue, la soddisfazione dei dipendenti verrà trattata come variabile quantitativa; se tali etichette si utilizzano per indicare un grado crescente di soddisfazione, la valutazione dei dipendenti continuerà ad essere qualitativa.

L'operazione di "quantificazione" di modalità ordinali mediante assegnazione di punteggi è dunque piuttosto arbitraria, e comunque poco generalizzabile; in questa tesi si tenterà, quindi, di applicare il metodo delle componenti principali mantenendo la natura ordinale delle variabili.

2.1 Il metodo di analisi delle componenti principali

Il metodo delle componenti principali consiste nel sostituire alle variabili originali, fra loro correlate, un numero inferiore di variabili ottenute come combinazioni lineari delle prime, con la caratteristica di essere tra loro non correlate e di spiegare una quota sufficiente dell'informazione globale contenuta nelle variabili iniziali. Così facendo, le nuove variabili permettono non solo di snellire la dimensione della matrice dei dati, ma possono essere usate in applicazioni successive, senza una significativa perdita di informazione.

Si consideri il vettore casuale $\mathbf{X}$ di m componenti:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix}$$

avente una certa distribuzione multivariata con vettore delle medie $\boldsymbol{\mu}$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix}$$

e matrice di varianze e covarianze $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \dots & \sigma_{mm} \end{pmatrix}$$

Al fine di rendere le variabili confrontabili, si standardizzano quelle originarie, trasformando il vettore $\mathbf{X}$ nel vettore $\mathbf{Z}$:

$$\mathbf{Z} = \begin{pmatrix} \dfrac{X_1 - \mu_1}{\sigma_1} \\ \dfrac{X_2 - \mu_2}{\sigma_2} \\ \vdots \\ \dfrac{X_m - \mu_m}{\sigma_m} \end{pmatrix}$$

Di conseguenza, la corrispondente matrice di varianze e covarianze di $\mathbf{Z}$ coinciderà con la matrice di correlazione:

$$\mathbf{P} = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1m} \\ \rho_{21} & 1 & \dots & \rho_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{m1} & \rho_{m2} & \dots & 1 \end{pmatrix}$$

dove $\rho_{ij} = \dfrac{cov(X_i, X_j)}{\sigma_i \sigma j}$.

Per la variabile casuale, combinazione lineare delle variabili standardizzate

$$Y_1 = \mathbf{a_1'Z} \tag{2.1}$$

occorre massimizzare la varianza

$$V(Y_1) = \mathbf{a_1'Pa_1} \tag{2.2}$$

sotto il vincolo di normalizzazione

$$\mathbf{a_1'a_1} = 1 \tag{2.3}$$

per mezzo del quale non si alterano le caratteristiche metriche dello spazio considerato.

La soluzione del problema si ottiene massimizzando il *Lagrangiano*:

$$L(\mathbf{a}, \lambda) = \mathbf{a_1'Pa_1} - \lambda(\mathbf{a_1'a_1} - 1) \tag{2.4}$$

$$\frac{\partial L(\mathbf{a}, \lambda)}{\partial \mathbf{a}} = 2(\mathbf{Pa_1} - \mathbf{a}\lambda) = 2(\mathbf{P} - \lambda\mathbf{I_m})\mathbf{a_1} = \mathbf{0_m} \tag{2.5}$$

$$(\mathbf{P} - \lambda\mathbf{I_m})\mathbf{a_1} = \mathbf{0_m} \tag{2.6}$$

Il sistema di equazioni omogeneo ammette come soluzione banale il vettore nullo. La condizione in base alla quale si ottengono soluzioni diverse dal vettore nullo è la seguente:

$$|\mathbf{P} - \lambda\mathbf{I_m}| = \mathbf{0_m} \tag{2.7}$$

Essendo questa un'equazione di grado m, la soluzione è data dagli m autovalori reali (distinti e non), $\lambda_1, \lambda_2, \ldots, \lambda_m$.

Per stabilire quale autovalore fornisce il massimo della funzione obiettivo, premoltiplichiamo per $\mathbf{a'}$ entrambi i membri della seguente espressione:

$$\mathbf{Pa_1} = \lambda\mathbf{a_1} \tag{2.8}$$

ottenendo:

$$\mathbf{a_1'Pa_1} = \mathbf{a_1'}\lambda\mathbf{a} \tag{2.9}$$

e poiché vale il vincolo di normalizzazione, la (2.9) diventa:

$$V(\mathbf{Y_1}) = \lambda \tag{2.10}$$

Poiché il nostro obiettivo è proprio massimizzare la varianza (2.10), la soluzione ottima si ricava in corrispondenza del massimo autovalore λ_1 della matrice di varianze e covarianze della variabile $\mathbf{Z}$.

La prima componente principale delle variabili Z_1, Z_2, ..., Z_m è, quindi, la combinazione lineare $Y_1 = a_{11}Z_1 + a_{21}Z_2 + \cdots + a_{m1}Z_m$ i cui coefficienti a_{i1} sono gli elementi del vettore caratteristico associato con la più grande radice caratteristica λ_1 della matrice di correlazione $\mathbf{P}$. Poiché gli a_{i1} soddisfano il vincolo di normalizzazione ($\mathbf{a_1'a_1} = 1$), la radice caratteristica λ_1 è interpetrabile come la varianza di Y_1.

La seconda componente principale Y_2 si ottiene come combinazione lineare di Z_1,Z_2,...,Z_m, caratterizzata da varianza massima ($V(Y_2) = \mathbf{a_2'Pa_2} = max$) sotto i vincoli di normalizzazione ($\mathbf{a_2'a_2} = 1$) e di incorrelazione con la precedente componente Y_1 ($\mathbf{a_2'Pa_1} = 0$),impostando un ulteriore problema di massimo vincolato:

$$L(\mathbf{a}, \lambda, \delta) = \mathbf{a_2'Pa_2} - 2\lambda\mathbf{a_2'a_2} + \lambda - \delta\mathbf{a_2'Pa_1} \tag{2.11}$$

$$\frac{\partial L(\mathbf{a}, \lambda, \delta)}{\partial \mathbf{a}} = 2\mathbf{Pa_2} - 2\lambda\mathbf{a_2} - \delta\mathbf{Pa_1} = 0 \tag{2.12}$$

La soluzione è fornita dall'autovettore di $\mathbf{P}$ corrispondente al secondo autovalore.

In definitiva, ripetendo questo procedimento fino ad Y_m si troveranno m nuove variabili casuali. Se le prime p (con $p < m$) di tali variabili forniscono una quota rilevante dell'informazione contenuta nelle variabili originali, possono essere usate in loro sostituzione.

Indicando con $\mathbf{A}$ la matrice degli autovettori di $\mathbf{P}$, soddisfacente il vincolo di orto-normalizzazione $\mathbf{A'A} = \mathbf{I}$, la trasformazione da m a m variabili corrisponde ad

una rotazione ortogonale degli assi cartesiani. I nuovi assi definiti dallo spazio degli autovettori sono detti "assi principali" e le nuove variabili Y_1, Y_2, ..., Y_m sono le "componenti principali" di $\mathbf{Z}$ (figura 2.1).

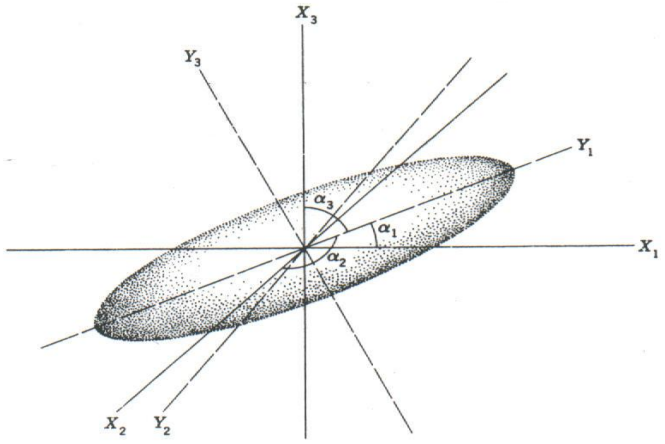

Figura 2.1: Rappresentazione grafica delle componenti principali

Per la scelta delle componenti principali si può fare riferimento agli autovalori con valore superiore ad uno, includendo solo le componenti che spiegano una varianza maggiore di quella introdotta nel modello da ogni singola variabile originaria. Un altro criterio di scelta è quello che tiene conto della varianza totale spiegata dalla i-esima componente, ossia $\frac{\lambda_i}{m}$. Se le prime p variabili spiegano una quota di varianza totale non inferiore ad una quantità prefissata, queste saranno le p componenti che si useranno come sintesi delle m variabili originarie.

2.2 Misure di associazione ordinale

Poiché le variabili misurate su scala ordinale non soddisfano gli assiomi delle misure metriche, non è sensato parlare di linearità. Tuttavia, l'ordinamento delle categorie permette di considerare il concetto di monotonicità.

Per applicare il metodo delle componenti principali a variabili ordinali, senza alterare l'informazione qualitativa dei dati, si può operare direttamente sulla matrice di correlazione, sostituendo i coefficienti di correlazione di Pearson con misure di associazione più adeguate, che descrivano il grado di monotonicità della relazione e che prendono il nome di *misure di associazione ordinale (o cograduazione).*

In generale due variabili *cograduano* se a valori alti dell'una corrispondono valori alti dell'altra (e a valori bassi dell'una, valori bassi dell'altra), *contro-graduano* se a valori alti dell'una corrispondono valori bassi dell'altra. Le diagonali della tabella di contingenza rappresentano il punto di riferimento per individuare l'esistenza o meno di un'eventuale relazione tra variabili ordinali, come mostra la seguente figura.

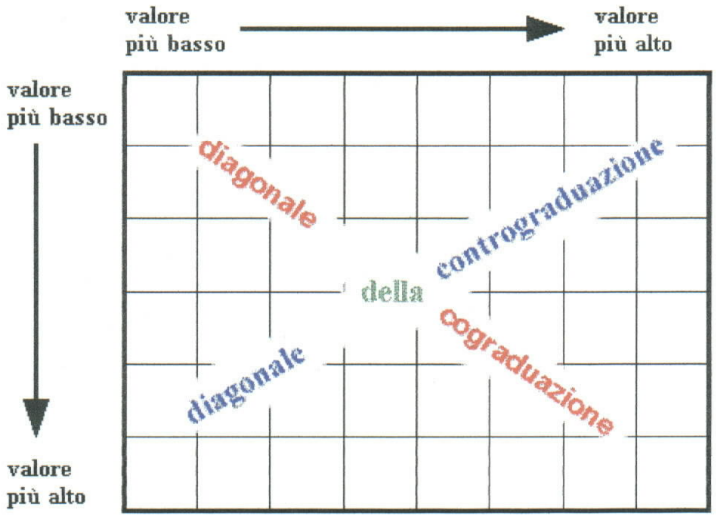

Figura 2.2: Le diagonali in una tabella di contingenza fra due variabili ordinali

Se le modalità delle due mutabili sono ordinate come in figura (2.2), cioè dal valore più basso al valore più alto, attribuendo per entrambe giudizi negativi ai valori più bassi e giudizi positivi a quelli più alti, la linea discendente che congiunge la cella "basso/basso" con la cella "alto/alto" è detta ***diagonale di cograduazione*** e la linea ascendente che congiunge la cella "alto in riga/basso in colonna" con la cella

"basso in riga/alto in colonna" è detta ***diagonale di contro-graduazione***. In realtà parlare di diagonali è improprio, dato che c'è una vera e propria diagonale solo se le due variabili hanno lo stesso numero di modalità. Naturalmente se le modalità sono sempre ordinate come in figura, ma la loro polarità è opposta, le diagonali di cograduazione e contro-graduazione si invertono.

In ogni caso fra due variabili ordinali c'è cograduazione se le frequenze della tabella di contingenza si addensano sulla (o attorno alla) diagonale di cograduazione; c'è contro-graduazione se le frequenze si addensano sulla (o attorno alla) diagonale di contro-graduazione.

Le misure di associazione ordinale più comunemente usate sono quelle basate sul numero di coppie *concordanti* (o cograduate) e *discordanti* (o contrograduate) delle osservazioni campionarie.

Una coppia di osservazioni (x_1, y_1) e (x_2, y_2) è concordante se l'ordinamento dei ranghi di x_1 e x_2 è uguale all'ordinamento dei ranghi di y_1 e y_2; discordante se l'ordinamento dei ranghi di x_1 e x_2 è opposto all'ordinamento dei ranghi di y_1 e y_2. Le coppie di osservazioni che non è possibile definire come concordanti o come discordanti, perché hanno la stessa modalità su una mutabile o su entrambe, si dicono *legate*.

Ad esempio: le coppie (2,3) e (1,2) sono concordanti, le coppie (2,3) e (1,4) sono discordanti, le coppie (2,3) e (2,5) sono legate sulla variabile di riga.

Data una tabella di contingenza le cui celle contengono le frequenze di tutte le combinazioni delle categorie delle due mutabili, generalmente indicate con n_{ij}, la formula generale per il calcolo di coppie concordanti e discordanti è la seguente:

$$C = \sum_{i<k} \sum_{j<l} n_{ij} n_{kl} \tag{2.13}$$

$$D = \sum_{i<k} \sum_{j>l} n_{ij} n_{kl} \tag{2.14}$$

In realtà non tutte le $\binom{n}{2}$ coppie di osservazioni sono concordanti o discordanti. In particolare:

- il numero totale di coppie legate sulla variabile di riga x è la somma del numero di coppie nella stessa riga, ossia:

$$T_x = \sum \frac{n_{i+}(n_{i+} - 1)}{2} \tag{2.15}$$

in cui n_{i+} è il numero di osservazioni nella riga i.

- il numero totale di coppie legate sulla variabile di colonna y è la somma del numero di coppie nella stessa colonna, cioé:

$$T_y = \sum \frac{n_{+j}(n_{+j} - 1)}{2} \tag{2.16}$$

in cui n_{+j} è il numero di osservazioni nella colonna j.

- il numero delle coppie legate sia su x che su y è dato dalla seguente formula:

$$T_{xy} = \sum_i \sum_j \frac{n_{ij}(n_{ij} - 1)}{2} \tag{2.17}$$

Poiché alcune delle coppie legate su x lo sono anche su y, al numero totale delle coppie (calcolato sommando quelle concordanti, quelle discordanti, quelle legate su x e quelle legate su y) bisogna sottrarre il numero delle coppie legate su x e su y. Si può quindi affermare che il numero totale delle coppie di osservazioni in una tabella di contingenza bidimensionale si può scomporre come segue:

$$\frac{n(n-1)}{2} = C + D + T_x + T_y - T_{xy} \tag{2.18}$$

nella quale n rappresenta il totale delle osservazioni campionarie.

I coefficienti più usati per quantificare il grado di cograduazione/contrograduazione sono: γ di Goodman e Kruskal, D di Somers, τ e τ_b di Kendall.

- **γ(Goodman, Kruskal, 1954)**

$$\hat{\gamma} = \frac{C - D}{C + D} \tag{2.19}$$

è il rapporto tra la differenza e la somma di coppie concordanti e discordanti. E' un coefficiente bi-direzionale e per $|\hat{\gamma}| = 1$ l'associazione è perfettamente monotona.

- **D (Somers, 1962)**

$$\hat{D}_{yx} = \frac{C - D}{\frac{n(n-1)}{2} - T_x} \tag{2.20}$$

$$\hat{D}_{xy} = \frac{C - D}{\frac{n(n-1)}{2} - T_y} \tag{2.21}$$

è la differenza tra le proporzioni di coppie concordanti e discordanti, eccetto quelle legate su x nella (2.20) o su y nella (2.21). Il coefficiente D di Somers è una misura da utilizzare quando tra le due variabili esiste una relazione asimmetrica.

- τ **(Kendall, 1938)**

$$\hat{\tau} = \frac{C - D}{\frac{n(n-1)}{2}} \tag{2.22}$$

è la differenza tra le proporzioni di coppie concordanti e discordanti, per tutte le coppie di osservazioni.

- $\tau_{\mathbf{b}}$ **(Kendall, 1945)**

$$\hat{\tau}_b = \frac{C - D}{\sqrt{(\frac{n(n-1)}{2} - T_x)(\frac{n(n-1)}{2} - T_y)}} \tag{2.23}$$

è la differenza tra le proporzioni di coppie concordanti e discordanti escluse le coppie di osservazioni legate su x e quelle legate su y.

Kendall ha proposto, quindi, due misure di associazione. Il coefficiente τ è stato introdotto nel 1938 per variabili continue, per le quali teoricamente non esistono coppie legate, cioè $T_x = T_y = T_{xy} = 0$. Nel caso di variabili categoriali, nelle quali è presente una grande proporzione di coppie legate, Kendall ha introdotto il coefficiente τ_b.

2.3 Il controllo sull'assenza di interazioni di ordine superiore al primo

Come già sottolineato nel par. 2.1, il metodo di analisi delle componenti principali considera un vettore casuale con una certa distribuzione multivariata e con vettore delle medie $\boldsymbol{\mu}$ e matrice di varianze e covarianze $\boldsymbol{\Sigma}$. Strettamente parlando, il metodo delle componenti principali è pienamente giustificato solo nei casi in cui il momento primo ed il momento secondo caratterizzano totalmente una distribuzione multivariata. In presenza di relazioni tra più di due variabili, l'informazione fornita dalle componenti principali risulta incompleta, dal momento che essa fa riferimento soltanto a misure di relazione bivariata. Diventa, quindi, importante verificare se esistono o meno interazioni di ordine superiore al primo.

Quando si opera con variabili categoriali, tale verifica può essere fatta confrontando modelli in cui si elimina (o si aggiunge) di volta in volta un parametro di interazione.

In linea teorica, poiché in questa tesi si concentra l'attenzione su variabili ordinali, sarebbe desiderabile realizzare questa verifica mediante modelli specifici per variabili ordinali. Tuttavia nell'applicazione presentata nel successivo cap. 3, per esaminare quali relazioni esistono tra le variabili ordinali, si ricorrerà all'utilizzo di modelli loglineari standard per variabili sconnesse.

Questa scelta è giustificata dal fatto che la conclusione cui si giunge con i modelli loglineari standard, quando si accetta l'ipotesi di assenza di interazione di ordine superiore al primo, non è diversa da quella cui si giunge utilizzando modelli loglineari per variabili ordinali. La differenza tra i due modelli, infatti, consiste non tanto sulla presenza o meno del fattore di interazione, quanto nel valore che questo assume.

Vi sono, inoltre, le seguenti ragioni pratiche per scegliere questo approccio:

- per tabelle multidimensionali ordinali non esistono generalizzazioni semplici universalmente accettate dei modelli loglineari ordinali sviluppati per tabelle

bidimensionali;

- modelli per variabili sconnesse possono sempre essere utilizzati anche per variabili ordinali, data l'esistenza della gerarchia tra le scale di misura;
- i software statistici a disposizione propongono soluzioni computazionalmente agevoli solo per il caso di dati categoriali sconnessi.

Per semplicità di notazione, si consideri una tabella di contingenza tridimensionale IxJxK in cui sono riportate le determinazioni campionarie delle tre variabili categoriali $X = \{x_1,\ldots,x_i,\ldots,x_I\}$, $Y = \{y_1,\ldots,y_j,\ldots,y_J\}$ e $Z = \{z_1,\ldots,z_k,\ldots,z_K\}$. Il *modello loglineare classico* si può definire come segue:

$$\log \mu_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \tag{2.24}$$

In base ad esso il logaritmo della frequenza attesa in una cella (i, j, k) si può esprimere come la somma di otto componenti: un effetto generale μ legato alla frequenza media generale, tre effetti marginali dovuti al fatto che la cella considerata appartiene alla categoria x_i di X, alla categoria y_j di Y, alla categoria z_k di Z, tre effetti di interazione di primo ordine tra variabili ed un effetto di interazione di secondo ordine tra la i-esima categoria di X, la j-esima categoria di Y e la k-esima categoria di Z.

I parametri sono sottoposti ai seguenti vincoli:

$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \ldots = \sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = \ldots = \sum_k \lambda_{ijk}^{XYZ} = 0$

ed il modello sopra riportato prende il nome di "modello loglineare saturo", in quanto il numero dei parametri è identico al numero delle celle della tabella di contingenza.

A partire da questo modello, tramite una procedura di *backward selection*, consistente nell'eliminazione di un termine parametrico alla volta, si possono individuare le interazioni significative. Se un parametro è non significativo, questo non verrà più incluso nel modello; in caso contrario, il parametro verrà reinserito nel modello. Le informazioni ricavabili dall'adattamento di un modello loglineare si possono

sintetizzare in tabelle strutturate in modo simile alle tabelle ANOVA per i modelli lineari. Nella prima colonna di ognuna di esse è riportata la descrizione del modello mediante la notazione "compatta" proposta da Upton (1978). In base ad essa, e in base al principio di gerarchia dei modelli utilizzati, la notazione identifica le interazioni di ordine massimo incluse nel corrispondente modello. Nella seconda colonna si riportano i valori della devianza residua; maggiore è tale valore e peggiore è l'adattamento del modello. I gradi di libertà, riportati nella terza colonna, evidenziano, crescendo, una maggiore semplicità del modello e quindi un minor numero di parametri presenti in esso. Nella quarta colonna si trova l' Akaike Information Criterion (AIC)(Akaike, 1973); questo è uno dei criteri di scelta, utilizzato per modelli nidificati, che si centra sul concetto di parsimonia: modelli più semplici sono preferiti a modelli più complessi. La sua formulazione, per ogni modello alternativo proposto per descrivere i dati, è la seguente:

$$AIC_i = -2logL(\hat{\theta}_i; \boldsymbol{n}) + 2(k - r_i) \tag{2.25}$$

in cui L($\hat{\theta}_i$;$\boldsymbol{n}$) è la funzione di verosimiglianza, $\boldsymbol{n}$ è il vettore delle frequenze della tabella di contingenza considerata e $(k - r_i)$ è il numero dei parametri presenti nell'iesimo modello; questa misura combina, quindi, bontà di adattamento e gradi di libertà. In una sequenza di modelli nidificati, si sceglie quello con l'AIC più basso.

Infine l'ultima colonna riporta il p-valore: più il suo valore è basso e più aumenta la probabilità di rigettare l'ipotesi H_0 di non significatività del parametro di interazione; quanto più il p-valore si avvicina all'unità, tanto più il valore della statistica calcolato nel campione assume valori bassi che indicano una forte plausibilità di H_0.

Se dall'analisi dei valori della suddetta tabella non si riscontrano interazioni di ordine superiore al primo e, quindi, tutte le relazioni tra variabili sono riassumibili in quelle a coppie, il metodo delle componenti principali sfrutterà tutta l'informazione disponibile e potrà essere applicato a variabili ordinali.

2.4 Analisi delle componenti principali su matrici di associazione ordinale

Il metodo di analisi delle componenti principali applicato a variabili categoriali ordinali tenterà di sostituire alle variabili ordinali originarie, fra loro cograduate, un numero inferiore di variabili ordinali, fra loro non cograduate. La matrice di cograduazione, che misura il grado di associazione tra le variabili originarie, dovrà essere simmetrica e semidefinita positiva; queste sono, infatti, le caratteristiche di una matrice di correlazione che stanno alla base dell'applicabilità del metodo delle componenti principali per variabili metriche.

La matrice di cograduazione è simmetrica, in quanto l'associazione tra due variabili A e B è uguale all' associazione tra B ed A. Inoltre, mediante la decomposizione spettrale di una matrice simmetrica $\mathcal{A}$, si dimostra che:

$$\lambda_i \geq 0 \qquad \forall i \Longleftrightarrow \mathcal{A} \text{ semidefinita positiva}$$

in cui i λ_i sono gli autovalori della matrice $\mathcal{A}$.

Se gli autovalori della matrice di cograduazione sono tutti positivi o nulli, essa è paragonabile alla matrice di correlazione; a questi autovalori è, quindi, possibile attribuire un significato di variabilità spiegata dalle componenti ordinali, parallelamente a quanto fatto in ambito metrico.

Individuata la matrice degli autovettori, sorge il problema di costruire le combinazioni lineari Y_1,Y_2,...,Y_m a partire da variabili che non hanno natura metrica.

A tal punto si rende indispensabile "quantificare" mediante punteggi numerici le modalità delle variabili ordinali per proseguire l'applicazione del metodo delle componenti principali. In una fase successiva, i valori corrispondenti alle componenti principali ottenute dovranno essere "ricategorizzati" per riacquisire la loro natura ordinale.

Ottenuto ciò, si potranno confrontare le combinazioni delle modalità delle mutabili originali con le categorie di risposta delle componenti principali, evidenziando

gli effetti dell'applicazione del metodo.

2.5 Gli impieghi

Uno degli impieghi più frequenti del metodo delle componenti principali riguarda la *formazione di graduatorie*.

Queste sono spesso utilizzate, ad esempio, nell'ambito della gestione delle risorse umane, per selezionare il personale sulla base dei risultati di test psico-attitudinali; oppure in ambito sociale, per la sintesi degli indicatori, o come si evidenzierà nell'applicazione successiva, per valutare la qualità dei corsi di una o più Facoltà sulla base della soddisfazione degli studenti.

Generalmente, per stilare tali graduatorie, si possono costruire indicatori aggregati; in alternativa, la formazione delle graduatorie può essere semplificata basandosi sulla sola prima componente principale, l'unica ad avere un significato di soddisfazione media, sempre che la porzione di varianza spiegata da essa non sia troppo bassa.

Tuttavia, come sarà evidenziato nell'applicazione basata sul questionario sulla valutazione della didattica universitaria nell'Ateneo di Siena, anche questo metodo presenta delle difficoltà. Dal momento che ogni corso è caratterizzato da un vettore di frequenze in corrispondenza delle nuove categorie ordinali individuate per la prima componente principale ordinale, sarà arduo stabilire un loro ordinamento totale, perché il confronto andrà fatto su più dimensioni, e spesso accade che le frequenze delle risposte alle modalità della prima componente principale, in corrispondenza ad ogni corso, non siano facilmente confrontabili.

Per risolvere queste difficoltà si potrà operare in modo diverso: prima di categorizzare i valori della prima componente principale, si calcoleranno le medie per ciascun corso; in questo modo la graduatoria dei corsi si potrà stilare sulla base del confronto tra scalari. Solo dopo aver "ricategorizzato" i valori medi, si potranno individuare i corsi con valutazione negativa, media o positiva.

Bibliografia

[1] Agresti A. (1996) *An introduction to categorical data analysis*. John Wiley, New York.

[2] Agresti A. (1984) *Analysis of ordinal categorical data*. John Wiley & Sons, New York.

[3] Morrison D. F. (1978) *Multivariate Statistical Methods*. International student edition, seconda edizione.

[4] Corbetta P. (1992) *Metodi di analisi multivariata per le scienze sociali*. Il Mulino, Bologna, parte seconda.

[5] Del Vecchio F. (1995) *Scale di misura e indicatori sociali*. Cacucci Editore, Bari.

[6] Everitt B. S. (1994) *The analysis of contingency tables*. John Wiley & Sons, New York.

[7] Kendall M. G. (1948) *Rank Correlation Methods*. Charles Griffin & Company, London.

[8] Kendall M. G., Gibbons G. D. (1990) *Rank Correlation Methods*. Oxford University, New York.

[9] http://www.cisi.unito.it/progetti/leda/UNITA19.HTM

[10] http://www.fire.ca.gov/BOF/poLfs/LewisHMP.pdf

[11] http://www.ms.uky.edu/˜rayens/teaching/sta673/chapter%205%20-Part%20ll.pdf

[12] http://www.cqs.washington.edu/papers/zabel/chp3.doc7.html

Appendice A

Il questionario per la valutazione della didattica nell'Ateneo di Siena

UNIVERSITÀ DEGLI STUDI DI SIENA

QUESTIONARIO DI VALUTAZIONE DELLA DIDATTICA

Istruzioni per la compilazione

1) Utilizzare esclusivamente PENNA NERA O BLU
2) Per ogni quesito annerire la casella corrispondente alla vostra risposta nel seguente modo [●]
3) Annerire una sola casella tra quelle previste per ogni quesito
4) In caso di errore cerchiare la casella errata (●) e annerire la casella corrispondente alla risposta corretta
5) Quando richiesto, si prega di scrivere le lettere (in stampatello) e i numeri nel modo più chiaro possibile, all'interno delle caselle predisposte

Esempio:

Cognome [R][O][S][S][I]

Codice [2][1][3]

Modalità di consegna

Il questionario deve essere inserito all'interno della busta fornita e restituito agli addetti alla distribuzione

Si ricorda che il questionario deve essere compilato **solamente una volta**

I questionari compilati sono **rigorosamente anonimi** e le informazioni in essi contenute saranno elaborate e diffuse solo in forma aggregata

GRAZIE PER LA COLLABORAZIONE

0000001165 A A

LO STUDENTE

FACOLTÀ E CORSO DI LAUREA ☐☐ (vedi legenda allegata)

A01 Sesso: ☐ maschio ☐ femmina ☐ non risponde

A02 Anno di nascita: 19 . .

terza cifra [0] [1] [2] [3] [4] [5] [6] [7] [8]

quarta cifra [0] [1] [2] [3] [4] [5] [6] [7] [8] [9]

A03 Scuola di provenienza:
☐ liceo classico ☐ ist. magistrale ☐ altri ist. tecnici
☐ liceo scientifico ☐ ist. tecnico commerciale ☐ ist. professionali
☐ altri licei ☐ ist. tecnico industriale ☐ altro

A04 Residenza:
☐ prov. Siena* ☐ altre regioni (nord) ☐ altre regioni (sud, isole)
☐ altra prov. toscana ☐ altre regioni (centro) ☐ Stato estero

A05 Durante il periodo delle lezioni: ☐ abita a Siena* ☐ è pendolare *

A06 Anno di corso:
In corso [1°] [2°] [3°] [4°] [5°] [6°]
Ripetente [1°] [2°] [3°] [4°] [5°] [6°]
Fuori corso [FC]

A07 Attività lavorativa nel periodo a cui si riferisce il questionario:
☐ nessuna ☐ saltuaria ☐ continuativa part-time ☐ continuativa a tempo pieno

A08 Numero di esami superati (escluse prove di idoneità):

prima cifra [0] [1] [2] [3]

seconda cifra [0] [1] [2] [3] [4] [5] [6] [7] [8] [9]

A09 Numero di corsi frequentati nel periodo (semestre o anno) a cui si riferisce il questionario:

[0] [1] [2] [3] [4] [5] [6] [7] [8] [9]

* Per la Facoltà di Lettere con sede in Arezzo, si consideri Arezzo

0000001165 B B

IL CORSO

(Attenzione: compilare tante schede di corso quanti sono i corsi indicati in A09)

B01 Codice del corso: ☐☐☐ (vedi legenda allegata)

Nome del corso (in stampatello): .....

B02 Con quale assiduità ha frequentato questo corso?

☐ *sempre o quasi (75% o più)* ☐ *abbastanza spesso (50-75%)* ☐ *saltuariamente (meno del 50%)*

Lo studente esprima il suo giudizio su ciascuno degli aspetti sotto indicati, graduando la risposta in una scala da 1 a 5 (1=per niente; 5=molto); utilizzare la casella** non si applica **quando l'aspetto in questione non è previsto nel corso

I CONTENUTI

C01 In che misura sono state fornite informazioni chiare ed esaurienti su:

- gli obiettivi del corso — *per niente* 1 2 3 4 5 *molto*
- il programma del corso — *per niente* 1 2 3 4 5 *molto*
- le modalità di svolgimento delle prove di esame — *per niente* 1 2 3 4 5 *molto*

C02 In che misura argomenti utili alla comprensione della materia sono stati trattati in modo lacunoso?

per niente 1 2 3 4 5 *molto*

Nel caso alcuni argomenti siano stati trattati in modo lacunoso, indichi almeno uno di tali argomenti

.....

.....

C03 In che misura gli argomenti del corso sono risultati una ripetizione inutile di argomenti trattati in altri corsi?

per niente 1 2 3 4 5 *molto*

Nel caso ci siano state ripetizioni inutili di argomenti già trattati, indichi almeno uno di tali argomenti

.....

.....

C04 In che misura i contenuti del corso si integrano con quelli degli altri insegnamenti?

per niente 1 2 3 4 5 *molto*

C05 In che misura la preparazione finora acquisita le ha permesso di seguire proficuamente il corso?

per niente 1 2 3 4 5 *molto*

C06 Come giudica la quantità di lavoro richiesta dal corso, anche in relazione agli altri insegnamenti?

per niente gravosa 1 2 3 4 5 *molto gravosa*

LE METODOLOGIE

C07 Come valuta, ai fini della comprensione della materia, la utilità di:

- lezioni formali — *per niente utile* 1 2 3 4 5 *molto utile* 0 *non si applica*
- seminari — *per niente utile* 1 2 3 4 5 *molto utile* 0 *non si applica*
- attività pratico-applicative (esercitazioni, lavoro per piccoli gruppi, altro) — *per niente utile* 1 2 3 4 5 *molto utile* 0 *non si applica*
- ricerche individuali/relazioni scritte — *per niente utile* 1 2 3 4 5 *molto utile* 0 *non si applica*

C C

LE RISORSE E L'ORGANIZZAZIONE

C08 **In che misura il materiale didattico fornito o indicato (testi, articoli, figure, bibliografie, etc) è risultato utile?**

per niente [1] [2] [3] [4] [5] *molto*

C09 **In che misura le attrezzature e gli strumenti utilizzati per le eventuali attività pratico/applicative sono risultati adeguati?**

per niente [1] [2] [3] [4] [5] *molto* [0] *non si applica*

C10 **In che misura la eventuale distribuzione di compiti di insegnamento tra più docenti all'interno del corso è risultata efficace?**

per niente [1] [2] [3] [4] [5] *molto* [0] *non si applica*

C11 **In che misura la frequenza del corso è risultata compatibile con quella degli altri corsi, in termini di:**

- orario *per niente* [1] [2] [3] [4] [5] *molto*
- localizzazione *per niente* [1] [2] [3] [4] [5] *molto*

SODDISFAZIONE COMPLESSIVA

C12 **Lo studente è pregato di indicare il suo livello di soddisfazione complessiva sul corso**

per niente soddisfatto [1] [2] [3] [4] [5] *molto soddisfatto*

IL DOCENTE / I DOCENTI DEL CORSO

Lo studente esprima il suo livello di soddisfazione sugli aspetti sotto indicati, graduando la risposta in una scala da 1 a 5 (1=per niente soddisfatto; 5=molto soddisfatto)

(Scrivere in stampatello all'interno delle caselle)

Docente 1 Cognome

Nome

D01 **Presenza regolare a lezione** [1] [2] [3] [4] [5]
D02 **Puntualità a lezione** [1] [2] [3] [4] [5]
D03 **Capacità espositiva** [1] [2] [3] [4] [5]
D04 **Capacità di suscitare interesse** [1] [2] [3] [4] [5]
D05 **Disponibilità a fornire chiarimenti** [1] [2] [3] [4] [5]
D06 **Reperibilità in orario di ricevimento** [1] [2] [3] [4] [5]

(Scrivere in stampatello all'interno delle caselle)

Docente 2 Cognome

Nome

D01 **Presenza regolare a lezione** [1] [2] [3] [4] [5]
D02 **Puntualità a lezione** [1] [2] [3] [4] [5]
D03 **Capacità espositiva** [1] [2] [3] [4] [5]
D04 **Capacità di suscitare interesse** [1] [2] [3] [4] [5]
D05 **Disponibilità a fornire chiarimenti** [1] [2] [3] [4] [5]
D06 **Reperibilità in orario di ricevimento** [1] [2] [3] [4] [5]